TITLE: Kbdock – Searching and Organising the Structural Space of Protein-Protein Interactions

AUTHORS: Marie-Dominique Devignes, Malika Smaïl-Tabbone, David Ritchie

TEASER: Big data is a recurring problem in structural bioinformatics where even just one experimentally determined protein structure can contain several different interacting protein domains and often involves many tens of thousands of 3D atomic coordinates. If we consider all protein structures that have ever been solved, the immense structural space of protein-protein interactions needs to be organised systematically in order to make sense of the many functional and evolutionary relationships that exist between different protein families and their interactions. This article describes some new developments in Kbdock, a knowledge-based approach for classifying and annotating protein interactions at the protein domain level.

Protein-protein interactions (PPIs) are fundamental biophysical interactions. Our understanding of many biological processes relies on the 3D-modelling of PPIs, and there is a growing need to be able to classify and analyse the structural repertoire of known PPIs using computational modelling and analysis techniques. Today, there exists over 105,000 experimentally determined 3D structures of proteins and protein complexes in the publicly available Protein Data Bank (PDB; http://www.rcsb.org/pdb/home/home.do). Each entry is composed of the 3D coordinates several thousands (sometimes even millions) of atoms belonging to one or more linear chains of amino-acid residues. By analysing multiple protein structures and their amino-acid sequences, it can be seen that many protein chains are modular, being composed of one or more structural domains. Thus, protein domains may be considered as "knowledge units" because they represent abstract classes which group together proteins with similar amino-acid sub-sequences and similar biological properties. We started to develop Kbdock (Kb for "knowledge-based") to address the problem of PPI modelling and classification at the domain level. In fact, as in many other complex scientific domains, Big Data approaches in the life sciences can only become viable by explicitly considering and making use of prior knowledge.

Essentially, Kbdock is a dedicated relational database which combines the Pfam domain classification (http://pfam.xfam.org/) with coordinate data from the PDB to analyse and model domain-domain interactions (DDIs) in 3D space. The detection of a DDI instance relies on the computation of the spatial distance between the surface residues of the domain instances found in each PDB entry. In the latest release of Kbdock, after duplicate or near-duplicate interactions are removed, a total of 5,139 distinct non-redundant DDIs involving two different domains have been identified from nearly 240,000 DDI instances extracted from the PDB. As illustrated in Fig 1A, the Kbdock resource can be queried by Pfam domain to visualise the DDI network involving this domain. Otherwise, Kbdock can return the list of DDI instances corresponding to a given pair of domains contained in two interacting proteins, even when there is little or no similarity with the query proteins. Moreover, calculating 3D super-positions of all DDI instances involving a given domain (Fig 1B) enabled us to perform a spatial clustering of all DDIs involving that domain, thereby identifying a discrete number of so-called "domain family binding sites" (DFBSs) [1] on the domain of interest. This gives us a new and original kind of knowledge unit, which we find to be essential when studying the structure and specificity of protein binding sites on a large scale [2]. The notion of DFBSs also led us to propose a case-based reasoning approach to the problem of how to retrieve the best available template for modeling protein-protein "docking" interactions.

The Kbdock project is run as a collaboration between the Capsid and Orpailleur teams at the Loria/Inria research center in Nancy. It is funded and supported by Inria, the CNRS, and the University of Lorraine, as well as specific ANR ("Agence Nationale pour la Recherche") grants. The Kbdock program is available through its on-line interface (http://kbdock.loria.fr/). It may also

be queried programmatically by expert users in order to execute complex or specialised queries. Recent developments to Kbdock make use of a novel protein structure alignment algorithm called “Kpax” (http://kpax.loria.fr) that we have developed [3]. This allows queries in Kbdock to span structural neighbours of the retrieved DDIs, thus allowing Kbdock to search over more distant regions of protein structure space and to propose protein docking templates that cannot be found using conventional sequence-based or structure-based comparison techniques.

We are currently working to link KBdock's structural domain binding site classification with the widely used ExPASy Enzyme Classification (http://enzyme.expasy.org) scheme. In order to achieve this, we are developing efficient data-mining approaches to process the millions of sequence-function associations that are now available in large molecular biology databases such as Swiss-Prot and TrEMBL, which together build the UniProt Knowledgebase (http://www.ebi.ac.uk/uniprot) at the European Bioinformatics Institute.

Useful Links

http://kbdock.loria.fr

http://kpax.loria.fr

http://hex.loria.fr

Références

1. Ghoorah AW, Devignes MD, Smaïl-Tabbone M, Ritchie DW. Spatial clustering of protein binding sites for template based protein docking. Bioinformatics. 2011 Oct 15;27(20):2820-7.
2. Ghoorah AW, Devignes MD, Alborzi SZ, Smaïl-Tabbone M, Ritchie DW. A structure-based classification and analysis of protein domain family binding sites and their interactions. Biology (Basel). 2015 Apr 9;4(2):327-43.
3. Ritchie DW, Ghoorah AW, Mavridis L, Venkatraman V. Fast protein structure alignment using Gaussian overlap scoring of backbone peptide fragment similarity. Bioinformatics. 2012 Dec 15;28(24):3274-81.

Contact address:

David Ritchie

Tel. +33 (0)3 83 59 30 45

E-mail dave.ritchie@inria.fr

Legend to the figure (see .jpg file ; a better resolution file can be provided soon)

Figure 1: Kbdock answers for a given domain (PF02943: FeThRed_B or Ferredoxin-Thioredoxin Reductase beta chain): (A) the graph of DDIs around this domain in Kbdock (depth = 2); (B) the superposed 3D DDI instances involving this domain.